CorReg : Prétraitement en régression linéaire par modélisation explicite des corrélations. Application aux valeurs manquantes

Clément Théry[1] & Christophe Biernacki[2] & Gaétan Loridant[3]

[1] *ArcelorMittal, Université Lille 1, Inria, CNRS, clement.thery@arcelormittal.com*
[2] *Université Lille 1, Inria, CNRS, christophe.biernacki@math.univ-lille1.fr*
[3] *Etudes Industrielles ArcelorMittal Dunkerque, gaetan.loridant@arcelormittal.com*

Résumé. La régression linéaire suppose en général l'usage de variables explicatives décorrélées, hypothèse souvent irréaliste pour les bases de données d'origine industrielle où les corrélations sont nombreuses et mènent à des estimateurs dégénérés. Le modèle proposé explicite les corrélations présentes sous la forme d'une famille de régressions linéaires entre covariables, permettant d'obtenir par marginalisation un modèle de régression parcimonieux libéré des corrélations, facilement interprétable et compatible avec les méthodes de sélection de variables. La structure de corrélations est estimée à l'aide d'un algorithme MCMC qui maximise la vraisemblance de la loi marginale sur les données. Un package R dénommé CorReg (sur le CRAN) permet la mise en oeuvre de cette méthode. Le modèle génératif sur les données et la modélisation explicite des corrélations permettent de gérer les valeurs manquantes, c'est cette conséquence de CorReg qui sera présentée.

Mots-clés. Régression, corrélations, industrie, valeurs manquantes, modèles génératifs

Abstract. Linear regression generally suppose to have decorrelated covariates. This hypothesis is often irrealist with industrial datasets that contains many highly correlated covariates leading to degenerated estimators. The proposed generative model consists in explicit modeling of the correlations with a family of linear regressions between the covariates permitting to obtain by marginalization a parsimonious correlation-free regression model, easily understandable and compatible with variable selection methods. The structure of correlations is found with an MCMC algorithm. An R package (CorReg) available on the CRAN implements this new method. We will describe missing values management, that is a consequence of the full generative model on the dataset and the explicit modeling of correlations.

Keywords. Regression, correlations, industry, missing values, generative models

1 Introduction

La régression linéaire classique suppose la décorrélation des covariables, source de problèmes en termes de variance des estimateurs. En effet, pour une variable réponse $\boldsymbol{Y} \in \mathcal{R}^n$ et un

ensemble de covariables $\boldsymbol{X} \in \mathcal{R}^{n\times d}$, la régression $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_Y$ avec $\boldsymbol{\varepsilon}_Y \sim \mathcal{N}(0, \sigma_Y^2 \boldsymbol{I}_n)$ (où $\boldsymbol{I}_n$ est la matrice identité de taille n) et $\boldsymbol{\beta} \in \mathcal{R}^d$ vecteur des d coefficients donne un estimateur $\hat{\boldsymbol{\beta}}$ de variance $\mathrm{Var}\,(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma_Y^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ dégénéré si les colonnes de $\boldsymbol{X}$ sont linéairement corrélées. Les méthodes de sélection comme le LASSO [4] muni du LAR [1] sont elles-mêmes touchées par ce problème de corrélation [5].

Notre idée est de modéliser explicitement les corrélations présentes entre covariables sous la forme d'une famille de régressions entre celles-ci. Nous présenterons donc le modèle génératif associé puis en partie 3 l'algorithme MCMC permettant d'estimer la famille de régressions à utiliser avant d'illustrer dans la partie 4 l'efficacité de la méthode sur des données simulées puis l'utilisation du modèle de sous-régression pour gérer les valeurs manquantes (partie 5) avant de conclure en partie 6.

2 Modèle supprimant les covariables corrélées

On suppose le modèle génératif suivant :

- Régression principale entre $\boldsymbol{Y}$ et $\boldsymbol{X}$:

$$\boldsymbol{Y}_{|\boldsymbol{X},\boldsymbol{S}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_Y = \boldsymbol{X}_f\boldsymbol{\beta}_f + \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}_Y \text{ avec } \boldsymbol{\varepsilon}_Y \sim \mathcal{N}(0, \sigma_Y^2\boldsymbol{I}_n); \tag{1}$$

- Famille de d_r régressions entre covariables de $\boldsymbol{X}$ corrélées :

$$\forall j \in \boldsymbol{J}_r : \ \boldsymbol{X}^{J_r^j}_{|\boldsymbol{X}_f,\boldsymbol{S}} = \boldsymbol{X}^{J_p^j}\boldsymbol{\alpha}_j + \boldsymbol{\varepsilon}_j \text{ avec } \boldsymbol{\varepsilon}_j \sim \mathcal{N}(0, \sigma_j^2\boldsymbol{I}_n); \tag{2}$$

- Mélanges gaussiens *indépendants* pour les covariables non corrélées :

$$\forall j \notin \boldsymbol{J}_r : \ \boldsymbol{X}^j \sim \sum_{k=1}^{k_j} \pi_k \mathcal{N}(\mu_{k_j}, \sigma_{k_j}^2\boldsymbol{I}_n); \tag{3}$$

où $\boldsymbol{J}_r$ est le vecteur des d_r indices des variables corrélées à gauche dans (2) et $\boldsymbol{J}_p = \{\boldsymbol{J}_p^1, \ldots, \boldsymbol{J}_p^{d_r}\}$ est le d_r-uplet des ensembles des indices des variables à droite dans (2). Les $\boldsymbol{\alpha}_j \in \mathcal{R}^{d_p^j}$ sont les coefficients des régressions entre covariables. On suppose que l'on a une partition des données $\boldsymbol{X} = (\boldsymbol{X}_r, \boldsymbol{X}_f)$ où $\boldsymbol{X}_r$ est la matrice des variables "redondantes" et $\boldsymbol{X}_f = \boldsymbol{X} \setminus \boldsymbol{X}_r$ la matrice complémentaire des variables indépendantes, *i.e.* les variables dépendantes dans $\boldsymbol{X}$ n'en expliquent pas d'autres. On note $d_r = \sharp\boldsymbol{J}_r$ le nombre de régressions entre covariables et $\boldsymbol{d}_p = (d_p^1, \ldots, d_p^{d_r})$ qui est le vecteur des longueurs des sous-régressions au sein de $\boldsymbol{X}$ avec $d_p^j = \sharp J_p^j$.

On a ainsi rendu explicites les corrélations au sein de $\boldsymbol{X}$ sous la forme d'une structure de sous-régressions linéaires $\boldsymbol{S} = (\boldsymbol{J}_r, \boldsymbol{J}_p)$.

On remarque alors que (1) et (2) impliquent par simple intégration sur $\boldsymbol{X}_r$, un modèle de régression en $\boldsymbol{Y}$ s'exprimant *uniquement en fonction des variables non corrélées* $\boldsymbol{X}_f$:

$$\boldsymbol{Y}_{|\boldsymbol{X}_f,\boldsymbol{S}} \quad = \quad \boldsymbol{X}_f(\boldsymbol{\beta}_f + \sum_{j=1}^{d_r} \beta_{J_r^j}\boldsymbol{\alpha}_j^*) + \boldsymbol{\varepsilon}\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}_Y = \boldsymbol{X}_f\boldsymbol{\beta}_f^* + \boldsymbol{\varepsilon}_Y^* \tag{4}$$

où $\boldsymbol{\alpha}_j^*$ est le vecteur $\boldsymbol{\alpha}_j$ complété par des 0 pour être de dimension $d - d_r$ (les sous-regressions sont supposées parcimonieuses), et $\boldsymbol{\varepsilon}$ est la matrice ayant les $\boldsymbol{\varepsilon}_j$ pour colonnes. L'estimateur classique du Maximum de Vraisemblance de $\boldsymbol{\beta_f}^*$ est sans biais et s'écrit

$$\hat{\boldsymbol{\beta}}_f^* = (\boldsymbol{X}'_f\boldsymbol{X}_f)^{-1}\boldsymbol{X}'_f\boldsymbol{Y} \tag{5}$$

En particulier sa matrice de variance

$$\text{Var}\,[\hat{\boldsymbol{\beta}}_f|\boldsymbol{X},\boldsymbol{S}] = (\sigma_Y^2 + \sum_{j\in I_2}\sigma_j^2\boldsymbol{\beta}_{J_r^j}^2)(\boldsymbol{X}'_f\boldsymbol{X}_f)^{-1} \tag{6}$$

peut être notablement mieux conditionnée que celle de $\hat{\boldsymbol{\beta}}$ initial (dimension réduite et surtout variables orthogonales). En outre, ce nouveau modèle réduit consiste en une régression linéaire classique qui peut donc bénéficier des outils de sélection de variables au même titre que le modèle complet. Notons enfin que la structure explicite entre les variables permet de mieux comprendre les phénomènes en jeu et la parcimonie du modèle facilite son interprétation.

Remarque : En ajoutant une étape de sélection de variables (de type LASSO) on obtient ainsi deux "types de 0" : ceux issus de l'étape de décorrélation et ceux issus de la sélection, avec deux interprétations différentes.

3 Estimation de la structure de corrélation

Pour comparer des structures de dimensions différentes, on compare les vraisemblances pénalisées de la la jointe de $\boldsymbol{X}$ par un critère de type BIC [3], noté BIC* et prenant comme loi *a priori* sur $\boldsymbol{S}$ une loi uniforme hiérarchique
$\mathbb{P}(\boldsymbol{S}) = \mathbb{P}(\boldsymbol{J_p}|\boldsymbol{d}_p, \boldsymbol{J}_r, d_r)\mathbb{P}(\boldsymbol{d}_p|\boldsymbol{J}_r, d_r)\mathbb{P}(\boldsymbol{J}_r|d_r)\mathbb{P}(d_r)$ plutôt qu'une loi uniforme simple. On a donc :

$$BIC^* \quad = \quad BIC + \ln(\mathbb{P}(\boldsymbol{S})). \tag{7}$$

L'équiprobabilité ainsi supposée des d_r et d_p^j vient pénaliser davantage la compléxité sous l'hypothèse $d_r < \frac{d}{2}$, hypothèse réaliste sur le nombre de régressions entre covariables. La recherche du meilleur $\boldsymbol{S}$ est combinatoire et un algorithme MCMC est utilisé par soucis d'efficacité et de flexibilité.

A chaque étape de l'algorithme, pour $\boldsymbol{S} \in \mathcal{S}$ (ensemble des structures réalisables) on définit un voisinage $\mathcal{V}_{\boldsymbol{S}}$ et ensuite la fonction de transition est guidée par BIC^* de la façon suivante :

$$\forall(\boldsymbol{S}, \tilde{\boldsymbol{S}}) \in \mathcal{S}^2 : P(\boldsymbol{S}, \tilde{\boldsymbol{S}}) \quad = \quad \mathbf{1}_{\{\tilde{\boldsymbol{S}} \in \mathcal{V}_{\boldsymbol{S}}\}} \frac{\exp(-\frac{1}{2} BIC^*(\tilde{\boldsymbol{S}}))}{\sum_{\boldsymbol{S}_l \in \mathcal{V}_{\boldsymbol{S}}} \exp(-\frac{1}{2} BIC^*(\boldsymbol{S}_l))}. \tag{8}$$

La chaîne de Markov ainsi constituée est ergodique dans un espace d'états finis et possède une unique loi stationnaire dont le mode correspond à la structure qui optimise BIC^*.

L'intialisation peut se faire en utilisant la matrice des corrélations et/ou la méthode du Graphical Lasso [2]. La grande dimension de l'espace parcouru rend préférable [8] (pour un temps de calcul égal) l'utilisation de multiples chaînes courtes plutôt qu'une seule très longue (permettant aussi la parallélisation).

En pratique, on commence par estimer pour chaque variable de $\boldsymbol{X}$ sa densité sous l'hypothèse d'un mélange gaussien (avec un package comme Rmixmod [6] par exemple). On peut alors calculer la loi jointe de $\boldsymbol{X}$ pour chaque structure réalisable rencontrée durant l'algorithme MCMC. Sans cette hypothèse générative supplémentaire sur $\boldsymbol{X}_f$, l'utilisation de BIC^* serait compromise. Notons cependant la souplesse de cette hypothèse due à la grande flexibilité des mélanges gaussiens [7].

4 Résultats sur données simulées

L'ensemble de la méthode à été programmé dans un package R dénommé CorReg. Pour les simulations présentées dans les tableaux 1 et 2, chacune des configurations à été simulée 100 fois. Les tableaux affichent le nombre de variables dépendantes trouvées ("bon gauche"), le nombre de variables jugées dépendantes à tort ("faux gauche") et les erreurs moyennes en prédiction (MSE) sur Y à partir d'échantillons de validation de 1 000 individus. Pour l'ensemble des simulations $d = 40$, $\sigma_Y = 10$, $\sigma_j = 0.001$ pour tout $j \in \{1, \ldots, d_r\}$, les $\boldsymbol{X}$ indépendants suivent des mélanges gaussiens à $\lambda = 5$ classes de moyenne selon une loi de Poisson de paramètre λ et d'écart-type λ. Les coefficients des $\boldsymbol{\alpha}_j$ suivent la même loi de Poisson mais avec un signe aléatoire. On cherche ici à se comparer à la méthode LASSO quand le vrai modèle est constitué de corrélations 2 à 2. CorReg a travaillé avec d_r et $\boldsymbol{d}_p$ libres et a utilisé `Rmixmod` pour estimer les densités dans $\boldsymbol{X}_f$.

Les tableaux 1 et 2 montrent que CorReg est équivalent au LASSO en l'absence de corrélations et le bat quand les corrélations sont fortes. On retrouve le phénomène attendu du LASSO moins impacté par les corrélations quand n grandit. On constate enfin la convergence asymptotique de CorReg vers le vrai modèle de régression.

On remarque que quand p_2 augmente le LASSO commence à se ressaisir car il y a de plus en plus de faux modèles proches du vrai en termes de prédiction donc le LASSO trouve des modèles inconsistants en interprétation mais relativement corrects en prédiction.

		Qualité de $\hat{\boldsymbol{S}}$		Qualité de prédiction (MSE)		
n	d_r	bon gauche	faux gauche	LASSO	CorReg $\hat{\boldsymbol{S}}$	CorReg vrai $\boldsymbol{S}$
30	16	8.48	4.88	3 511 185.23	10 686.62	738.89
30	32	16.89	2.78	565.51	189.54	139.24
50	0	0	0	529.94	529.94	529.94
50	16	8.89	5.4	347.59	233.99	197.95
50	32	18.95	2.44	163.7	139.39	121.56
400	32	23.49	1.06	104.52	103.6	102.67

Table 1: $\boldsymbol{Y}$ dépend de $\boldsymbol{X}$ entier. CorReg gagne logiquement.

		Qualité de $\hat{\boldsymbol{S}}$		Qualité de prédiction (MSE)		
n	d_r	bon gauche	faux gauche	LASSO	CorReg $\hat{\boldsymbol{S}}$	CorReg vrai $\boldsymbol{S}$
30	16	8.29	5	5 851.45	559.58	340.29
30	32	17	2.59	893	196.01	135.78
50	16	8.98	5.19	201.56	164.58	162.49
50	32	19.05	2.32	172.93	136.77	121.19
400	32	23.51	1.09	104.49	103.02	102.26

Table 2: $\boldsymbol{Y}$ dépend de $\boldsymbol{X}_r$ uniquement (cas normalement défavorable à CorReg).

5 Application aux problèmes de valeurs manquantes

Un coproduit du modèle de sous-régression concerne les valeurs manquantes. Le fait de disposer d'un modèle génératif complet sur $\boldsymbol{X}$ avec modélisation explicite des dépendances permet en effet de composer avec les valeurs manquantes en utilisant les lois conditionnelles. Tout d'abord, l'estimation de $\boldsymbol{\alpha}$ peut se faire sur les données observées en intégrant sur les données manquantes. On peut alors utiliser un algorithme de type EM (Expectation Maximization) pour estimer $\hat{\boldsymbol{\alpha}}$.

En pratique, on fait appel à une variante de EM : l'algorithme Stochastic EM qui remplace l'étape E par une étape stochastique d'imputation des valeurs manquantes, par exemple en utilisant un échantillonneur de Gibbs. Cet algorithme de Gibbs peut alors être utilisé pour faire de l'imputation multiple sur les valeurs manquantes en s'appuyant sur le $\hat{\boldsymbol{\alpha}}$ issu du Stochastic EM. Comme cette imputation tient compte des corrélations entre les variables, elle est plus précise qu'une simple imputation par la moyenne. Un avantage de l'imputation multiple est que l'on peut avoir une idée de la robustesse des imputations en regardant simplement la variance des valeurs imputées. Encore une fois, on y gagne en qualité d'interprétation. Autrement dit, le modèle génératif sur $\boldsymbol{X}$ donne la loi conditionnelle des valeurs manquantes sachant les valeurs observées, ce qui permet

d'imputer les valeurs manquantes en connaissant la variance associée à ces imputations.

6 Conclusion et perspectives

CorReg est disponible sur le CRAN et a d'ores et déjà montré son efficacité sur de vraies problématiques de régression en entreprise. Sa force est la grande interprétabilité du modèle proposé, qui est constitué de plusieurs régression linéaires courtes et donc facilement accessibles aux non statisticiens tout en luttant efficacement contre les problématiques de corrélations, omniprésentes dans l'industrie. On note également la possibilité d'élargir le champ d'application à la gestion des valeurs manquantes, aussi très présentes dans l'industrie. Le modèle génératif actuel et la modélisation explicite des liaisons entre covariables permer de gérer les problématiques de valeurs manquantes et vient renforcer encore l'intérêt du modèle de sous-régressions. Enfin, le principe de CorReg qui est l'explicitation des régressions latentes entre covariables pourrait être appliqué à d'autres méthodes prédictives (régression logistique, *etc.*).

Bibliographie

[1] Efron, B., Hastie, T., Johnstone,I. et Tibshirani, R. (2004), Least angle regression. *The Annals of statistics*, 32(2):407-499.
[2] Friedman, J., Hastie, T. et Tibshirani, R. (2008), Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432-441 .
[3] Lebarbier, E. et Mary-Huard,T. (2006), Une introduction au critère bic: fondements théoriques et interprétation. *Journal de la SFdS* , 147(1):39-57.
[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society.* Series B (Methodological), pages 267-288.
[5] Zhao,P. et Yu,B. (2006), On model selection consistency of lasso, *J. Mach. Learn. Res.* 7:2541-2563.
[6] Biernacki, C., Celeux, G., Govaert, G., et Langrognet, F. (2006), Model-based cluster and discriminant analysis with the MIXMOD software, Computational Statistics & Data Analysis, 51(2), 587-600.
[7] McLachlan, G., et Peel, D. (2004). Finite mixture models. Wiley. com.
[8] Gilks, W. R., Richardson, S., et Spiegelhalter, D. J. (Eds.). (1996). Markov chain Monte Carlo in practice (Vol. 2). CRC press.